(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6520108号
(P6520108)
(45)発行日　令和1年5月29日(2019.5.29)　(24)登録日　令和1年5月10日(2019.5.10)

(51)Int.Cl.　F I
G10L 13/10 (2013.01)　G10L 13/10 112E
G10L 13/10 114

請求項の数 10　(全 14 頁)

(21)出願番号　特願2014-259485(P2014-259485)
(22)出願日　平成26年12月22日(2014.12.22)
(65)公開番号　特開2016-118722(P2016-118722A)
(43)公開日　平成28年6月30日(2016.6.30)
審査請求日　平成29年12月19日(2017.12.19)

(73)特許権者　000001443
カシオ計算機株式会社
東京都渋谷区本町１丁目６番２号
(74)代理人　100074099
弁理士　大菅　義之
(72)発明者　田中　飛雄太
東京都羽村市栄町３丁目２番１号　カシオ計算機株式会社羽村技術センター内

審査官　上田　雄

最終頁に続く

(54)【発明の名称】音声合成装置、方法、およびプログラム

(57)【特許請求の範囲】
【請求項１】
テキストデータに基づいて、音声素片を選択する選択部と、
前記選択された音声素片を接続することにより接続音声素片を生成する接続部と、
音声データから得られるピッチ列を量子化し、前記量子化されたピッチ列を平滑化し、前記平滑化されたピッチ列から抑揚情報を抽出する抑揚情報抽出部と、
前記接続音声素片に含まれる要素の少なくとも一部を前記抑揚情報に基づいて変更して合成音声を出力する出力部と、
を備える音声合成装置。
【請求項２】
前記出力部は、前記接続音声素片に含まれる要素であるピッチ列を、前記抑揚情報抽出部から抽出された抑揚情報に適応するように変更するピッチ適応部を含む、請求項１に記載の音声合成装置。
【請求項３】
前記平滑化されたピッチ列は、加重移動平均演算することにより平滑化されている、請求項１または２に記載の音声合成装置。
【請求項４】
前記ピッチ適応部はさらに、前記音声データに含まれる韻律情報としてのピッチ列と前記接続音声素片に含まれるピッチ列との時間スケールを調整するとともに、前記韻律情報としてのピッチ列と前記接続音声素片に含まれるピッチ列とのピッチ存在区間を調整する

、請求項２に記載の音声合成装置。
【請求項５】
前記出力部は、前記接続音声素片に含まれる要素であるパワー列を、前記抑揚情報抽出部から抽出された抑揚情報に適応するように変更するパワー適応部を含む、請求項１に記載の音声合成装置。
【請求項６】
前記抑揚情報抽出部は、前記音声データに含まれる韻律情報としてのパワー列を平滑化し、当該平滑化されたパワー列を前記抑揚情報として抽出し、
前記パワー適応部は、前記接続音声素片に含まれるパワー列を平滑化し、当該平滑化されたパワー列と前記抑揚情報としての平滑化されたパワー列との比の列を算出し、当該比の列に基づいて前記接続音声素片のパワー列を修正する、請求項５に記載の音声合成装置。
【請求項７】
前記平滑化されたパワー列は、前記パワー列に含まれるパワーそれぞれを加重移動平均演算することにより取得する、請求項６に記載の音声合成装置。
【請求項８】
前記パワー適応部はさらに、前記音声データに含まれる韻律情報としてのパワー列及び前記接続音声素片に含まれるパワー列それぞれの時間スケールを調整する、請求項５ないし７のいずれかに記載の音声合成装置。
【請求項９】
音声合成装置に用いられる音声合成方法であって、前記音声合成装置が、
テキストデータに基づいて、音声素片を選択し、
前記選択された音声素片を接続することにより接続音声素片を生成し、
音声データから得られるピッチ列を量子化し、前記量子化されたピッチ列を平滑化し、前記平滑化されたピッチ列から抑揚情報を抽出し、
前記接続音声素片に含まれる要素の少なくとも一部を前記抑揚情報に基づいて変更して合成音声を出力する、音声合成方法。
【請求項１０】
音声合成装置に用いられる音声合成装置として用いられるコンピュータに、
テキストデータに基づいて、音声素片を選択するステップと、
前記選択された音声素片を接続することにより接続音声素片を生成するステップと、
音声データから得られるピッチ列を量子化し、前記量子化されたピッチ列を平滑化し、前記平滑化されたピッチ列から抑揚情報を抽出するステップと、
前記接続音声素片に含まれる要素の少なくとも一部を前記抑揚情報に基づいて変更して合成音声を出力するステップと、
を実行させるプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声コーパスからの音声素片の選択によって音声合成を行う技術に関する。
【背景技術】
【０００２】
入力テキストデータから生成される合成目標に対して、電子化された大規模な言語・音声データである音声コーパスを参照することにより音声波形の素片（以下、「音声素片」と記載する）を選択し、当該音声素片を接続することにより合成音声を出力する音声合成技術が知られている（例えば非特許文献１～３に記載の技術）。
【０００３】
このような音声合成技術において、音声コーパスから合成目標に最も適合する音声素片列を選択するための手法として従来、次のような技術が知られている（例えば非特許文献３に記載の技術）。まず、入力テキストデータから抽出される音素列ごとに、その音素列

と同じ音素列を有する音声素片のデータ（以下、「素片データ」と記載する）が、素片候補データとして音声コーパスから抽出される。次に、ＤＰ（Ｄｙｎａｍｉｃ　Ｐｒｏｇｒａｍｍｉｎｇ：動的計画法）アルゴリズムによって、入力テキストデータ全体に渡ってコストが最小となる最良の素片候補データの組（最良の素片データ列）が決定される。コストとしては、入力テキストデータと音声コーパス内の各素片データ間の音素列および韻律の差異、素片候補データである隣接する素片データ間のスペクトラム包絡などの音響パラメータ（特徴量ベクトルデータ）の不連続性などが用いられる。

【０００４】

入力テキストデータに対応する音素列は、例えば入力テキストデータに対して形態素解析処理を実行することで得られる。

【０００５】

入力テキストデータに対応する韻律（以下これを「目標韻律」と記載する）は、音素ごとの声帯の基本周波数であるピッチの高さ、持続時間長、および強度（パワー）である。この目標韻律の指定方式としては、入力テキストデータから得られる言語情報をもとに、実際の音声データに基づく統計的なモデルを用いて生成する方法がある（例えば非特許文献４に記載の技術）。言語情報は、例えば入力テキストデータに対して形態素解析処理を実行することで得られる。あるいは、目標韻律の指定方式として、ユーザが数値でパラメータ入力する方法がある。

【０００６】

さらに、目標韻律の指定方式として、ユーザが自身でテキストを発声するなどして与える音声によって指定するという方法がある。この方式は、テキストからの推定や数値パラメータの調整と比較して直感的な操作が可能であり、感情や抑揚の付与など自由度の高い目標韻律指定が行えるという利点がある。

【先行技術文献】

【非特許文献】

【０００７】

【非特許文献１】“CHATR:自然音声波形接続型任意音声合成システム”、電子情報通信学会信学技法、SP96-7.

【非特許文献２】“大規模コーパスを用いた音声合成システムXIMERA”、電子情報通信学会論文誌D Vol.J89-D No.12 pp.2688-2698.

【非特許文献３】河井　恒、“知識ベース　3-4 コーパスベース音声合成”、［online］、ver.1/2011.1.7、電子情報通信学会、［平成２６年１２月５日検索］、インターネット＜ＵＲＬ：http://27.34.144.197/files/02/02gun_07hen_03.pdf#page=6＞

【非特許文献４】匂坂　芳典、“知識ベース　3-7 韻律の生成”、［online］、ver.1/2011.1.7、電子情報通信学会、［平成２６年１２月５日検索］、インターネット＜ＵＲＬ：http://27.34.144.197/files/02/02gun_07hen_03.pdf#page=13＞

【発明の概要】

【発明が解決しようとする課題】

【０００８】

しかし、ユーザの音声入力による目標韻律指定方式には以下の様な課題がある。まず、指定できる目標韻律の自由度が増すため、それに対応できるだけの音声素片が必要となり、十分な量を揃えようとすると音声コーパスのデータベースが巨大になってしまう。また、ユーザが入力した音声の目標韻律と音声データベース内の音声素片が持っている韻律とでは、例えば声の高さなど個人に依存する相違点があり、適切な音声素片を選択することが困難になる。

【０００９】

上記の課題を解決するために、音声波形接続処理時に信号処理によって、音声素片の以下の要素を補正し、ユーザの入力した音声の目標韻律に適応させる方法が知られている。

１．各音素の継続時間長。

２．ピッチ（音の高低）。

３．パワー（音の大小）。
しかし、信号処理によってユーザが入力した音声の目標韻律を音声データベースから選択した音声素片に単純に適応するだけでは、以下の様な問題が生じる。ユーザが入力した音声の目標韻律には細かなピッチやパワーの変動が含まれていることがあり、それら全てを音声素片に適応させると信号処理による音質の劣化が顕著になってしまう。また、ユーザの入力した音声の目標韻律と音声素片の韻律（特にピッチ）が大きく異なる場合、単純に適応してしまうと合成音声の音質が劣化してしまう。

【００１０】
そこで、本発明は、波形接続方式の音声合成システムにおいて目標韻律を音声入力によって指定する際に、音声コーパスの規模拡大させる必要なく、高い自由度を維持しつつ合成音声の音質を向上させることを目的とする。

【課題を解決するための手段】

【００１１】
態様の一例では、入力されたテキストデータに基づいて、音声素片を選択する選択部と、選択された音声素片を接続することにより接続音声素片を生成する接続部と、入力された音声データに含まれる韻律情報から抑揚情報を抽出する抑揚情報抽出部と、接続音声素片に含まれる要素の少なくとも一部を抑揚情報に基づいて変更して合成音声を出力する出力部と、を備える。

【発明の効果】

【００１２】
本発明によれば、波形接続方式の音声合成システムにおいて目標韻律を音声入力によって指定する際に、音声コーパスの規模拡大させる必要なく、高い自由度を維持しつつ合成音声の音質を向上させることが可能となる。

【図面の簡単な説明】

【００１３】
【図１】音声合成装置の実施形態のブロック図である。
【図２】音声ＤＢのデータ構成例を示す図である。
【図３】音声合成装置の実施形態のハードウェア構成例を示す図である。
【図４】音声合成処理の例を示すフローチャートである。
【図５】ピッチ適応処理の説明図である。
【図６】パワー適応処理の説明図である。
【図７】ピッチ適応処理の詳細例を示すフローチャートである。
【図８】パワー適応処理の詳細例を示すフローチャートである。

【発明を実施するための形態】

【００１４】
以下、本発明を実施するための形態について図面を参照しながら詳細に説明する。図１は、音声合成装置の実施形態１００のブロック図である。音声合成装置１００は、音声合成部１０１、音声データベース（以下「音声ＤＢ」と記載）１０２、入力部１０３、および出力部１０４を備える。さらに、音声合成部１０１は、テキスト解析モジュール１０５、韻律解析モジュール１０６、素片選択モジュール１０７、波形接続モジュール１０８、ピッチ適応モジュール１０９、パワー適応モジュール１１０、およびシステム制御部１１１を備える。また、入力部１０３は、音声入力装置１１２およびテキスト入力装置１１３を備える。出力部１０４は、音声出力装置１１４を備える。素片選択モジュール１０７および波形接続モジュール１０８は音声素片選択・接続部に対応し、ピッチ適応モジュール１０９およびパワー適応モジュール１１０は抑揚情報抽出部および抑揚適応部に対応する。

【００１５】
入力部１０３のテキスト入力装置１１３は、入力テキストデータを入力する。また、入力部１０３の音声入力装置１１２は、入力音声データを入力する。

【００１６】

音声合成部１０１は、テキスト入力装置１１３から入力される入力テキストデータから生成される合成目標に対して、音声ＤＢ１０２に記憶されている音声素片の集合である音声コーパスを参照することにより音声素片を選択し、その音声素片を接続することにより接続音声素片を生成する。

【００１７】

図２は、図１の音声ＤＢ１０２に記憶される音声コーパスのデータ構成例を示す図である。音声コーパスとしては、例えば下記項目のデータが格納されている。

・予め録音された音声データ（図２（ａ））。

・図２（ａ）の音声データに対する音素ラベルの情報（図２（ｂ））。基本的にこのラベル付けされた図２（ａ）の音声データの断片が音声素片となる。この音素ラベルの情報は、図２（ｂ）に示されるように、「開始位置」、「継続時間長」、および「音素種類」の各情報を有する。

・一定時間Ｔ（ｍｓ：ミリ秒）からなるセグメントごとに、図２（ａ）の音声データから解析されたピッチ、パワー、フォルマントなどの音響情報（図２（ｃ））。セグメント長Ｔは、例えば「１０」ｍｓである。

【００１８】

図１の説明に戻り、音声合成部１０１内のテキスト解析モジュール１０５は、テキスト入力装置１１３が入力した入力テキストデータに対して例えば形態素解析処理を実行することにより、入力テキストデータに対応するアクセント付きの音素列を抽出する。

【００１９】

音声合成部１０１内の韻律解析モジュール１０６は、音声入力装置１１２が入力した入力音声データを解析して、目標韻律を抽出する。

【００２０】

音声合成部１０１内の素片選択モジュール（音声素片選択・接続部）１０７は、入力テキストデータから生成された音素列と入力音声データから生成された目標韻律とからなる合成目標に対して、音声データ内の音声コーパス（図２）を参照することにより音声素片を選択する。

【００２１】

音声合成部１０１内の波形接続モジュール１０８は、素片選択モジュール１０７で選択された音声素片を接続することにより、接続音声素片を生成する。

【００２２】

音声合成部１０１内のピッチ適応モジュール１０９は、波形接続モジュール１０８が出力する接続音声素片に含まれるピッチ列を、入力部１０３の音声入力装置１１２から入力された入力音声データに含まれるピッチ列に適応させるように変更する。

【００２３】

音声合成部１０１内のパワー適応モジュール１１０は、波形接続モジュール１０８が出力する接続音声素片に含まれるパワー列を、入力部１０３の音声入力装置１１２から入力された入力音声データに含まれるパワー列に適応させるように変更する。

【００２４】

音声合成部１０１内のシステム制御部１１１は、音声合成部１０１内の１０５～１１０の各部分の動作の実行順序等を制御する。

【００２５】

図３は、図１の音声合成装置１００をソフトウェア処理として実現できるコンピュータのハードウェア構成例を示す図である。図３に示されるコンピュータは、ＣＰＵ３０１、ＲＯＭ（リードオンリーメモリ：読出し専用メモリ）３０２、ＲＡＭ（ランダムアクセスメモリ）３０３、入力装置３０４、出力装置３０５、外部記憶装置３０６、可搬記録媒体４１０が挿入される可搬記録媒体駆動装置３０７、及び通信インターフェース３０８を有し、これらがバス３０９によって相互に接続された構成を有する。同図に示される構成は上記システムを実現できるコンピュータの一例であり、そのようなコンピュータはこの構成に限定されるものではない。

【００２６】
ＲＯＭ３０２は、コンピュータを制御する音声合成プログラムを含む各プログラムを記憶するメモリである。ＲＡＭ３０３は、各プログラムの実行時に、ＲＯＭ３０２に記憶されているプログラム又はデータを一時的に格納するメモリである。
【００２７】
外部記憶装置３０６は、例えばＳＳＤ（ソリッドステートドライブ）記憶装置またはハードディスク記憶装置であり、入力テキストデータ、入力音声データ、接続音声素片データ、または合成音声データ等の保存に用いられる。また、外部記憶装置３０６は、図２のデータ構成例を有する音声コーパスが格納された音声ＤＢ１０２を記憶する。
【００２８】
ＣＰＵ３０１は、各プログラムを、ＲＯＭ３０２からＲＡＭ３０３に読み出して実行することにより、当該コンピュータ全体の制御を行う。
【００２９】
入力装置３０４は、ユーザによるキーボードやマウス等による入力操作を検出し、その検出結果をＣＰＵ３０１に通知する。また、入力装置３０４は、図１の入力部１０３の音声入力装置１１２の機能を備え、特には図示しないマイクまたはライン入力端子を介して入力音声データを入力し、Ａ／Ｄ（アナログ－デジタル）変換によりデジタルデータに変換した後に、ＲＡＭ３０３または外部記憶装置３０６に記憶させる。さらに、入力装置３０４は、図１の入力部１０３のテキスト入力装置１１３の機能を備え、特には図示しないキーボードまたはデバイスインタフェース等を介して入力テキストデータを入力し、ＲＡＭ３０３または外部記憶装置３０６に記憶させる。
【００３０】
出力装置３０５は、ＣＰＵ３０１の制御によって送られてくるデータを表示装置や印刷装置に出力する。また、出力装置３０５は、ＣＰＵ３０１が外部記憶装置３０６またはＲＡＭ３０３に出力した合成音声データを、特には図示しないが、Ｄ／Ａ変換器でアナログ合成音声信号に変換した後、増幅器で増幅し、スピーカを介して、合成音声として放音する。
【００３１】
可搬記録媒体駆動装置３０７は、光ディスクやＳＤＲＡＭ、コンパクトフラッシュ等の可搬記録媒体３１０を収容するもので、外部記憶装置３０６の補助の役割を有する。
【００３２】
通信インターフェース３０８は、例えばＬＡＮ（ローカルエリアネットワーク）又はＷＡＮ（ワイドエリアネットワーク）の通信回線を接続するための装置である。
【００３３】
本実施形態による音声合成装置１００では、ＣＰＵ３０１が、ＲＯＭ３０２に記憶された音声合成プログラムを、ＲＡＭ３０３をワークメモリとして使用しながら実行することにより、図１の音声合成部１０１内の１０５～１１１の各ブロックの機能を実現する。そのプログラムは、例えば外部記憶装置３０６や可搬記録媒体４１０に記録して配布してもよく、或いはネットワーク接続装置３０８によりネットワークから取得できるようにしてもよい。
【００３４】
図４は、図１の構成に対応する音声合成装置１００の機能を、図３のハードウェア構成例を有するコンピュータのＣＰＵ３０１が、ソフトウェアプログラムの処理により実現する場合の、音声合成処理の例を示すフローチャートである。以下随時、図１、図２、および図３を参照する。
【００３５】
ＣＰＵ３０１はまず、テキスト入力装置１１３が入力した入力テキストデータに対して、テキスト解析処理を実行する（ステップＳ４０１）。ここでは、ＣＰＵ３０１は、入力テキストデータに対して例えば例えば形態素解析処理を実行することにより、入力テキストデータに対応するアクセント付きの音素列を抽出する。この処理は、図１のテキスト解

析モジュール１０５の機能を実現する。

【００３６】

次に、ＣＰＵ３０１は、音声入力装置１１２が入力した入力音声データに対して、韻律解析処理を実行する（ステップＳ４０２）。ここでは、ＣＰＵ３０１は、入力音声データに対して例えばピッチ抽出処理とパワー分析処理を実行する。そして、ＣＰＵ３０１は、ステップＳ４０２のテキスト解析処理により得られたアクセント付きの音素列を参照することにより、音素ごとのピッチの高さ（周波数）、持続時間長、およびパワー（強度）を算出し、それらの情報を目標韻律として出力する。

【００３７】

次に、ＣＰＵ３０１は、素片選択処理を実行する（ステップ４０３）。ここでは、ＣＰＵ３０１は、図２に例示されるデータ構成を有する音声コーパスが登録されている音声ＤＢ１０２から、音素および韻律に関して計算されるコストが最良になるように、ステップＳ４０１で算出された音素列およびステップＳ４０２で算出された目標韻律に対応する音声素片の列を選択する。このとき、ＣＰＵ３０１はまず、音声コーパス中の音素ラベルの列（図２（ｂ））をステップＳ４０１で算出された音素列と比較することにより、素片評価のコスト条件を満たす素片候補データを音声コーパスからリストアップする。次に、ＣＰＵ３０１は、素片候補データにおける音響情報（図２（ｃ））を目標韻律と比較することにより、接続評価のコスト条件を満たす最良の素片候補データを、リストアップした素片候補データから選択し、最終的に、音声素片の列を選定する。

【００３８】

次に、ＣＰＵ３０１は、波形接続処理を実行する（ステップＳ４０４）。ここでは、ＣＰＵ３０１は、ステップＳ４０３での音声素片の選択結果を入力して、対応する音声素片の音声データ（図２（ａ））を音声ＤＢ１０２中の音声コーパスから抽出し、それらを接続して接続音声素片を出力する。

【００３９】

上述のようにして出力された接続音声素片は、音声ＤＢ１０２が保有している音声コーパス内で、入力された音素列と目標韻律に対して、音素に関する素片評価と韻律に関する接続評価を合わせたコストが最良になるように選択されたものである。しかし、音声コーパスとして巨大なデータベースを保有できないような小規模のシステムにおいては、入力音声データから生成された目標韻律と音声コーパス内の限られた規模の音声素片が持っている韻律とでは、抑揚の付け方などに関して個人に依存する相違点がある。このため、ステップＳ４０４で接続音声素片が出力された段階では、入力音声データで表現されている抑揚が接続音声素片に十分に反映されているとはいえない。一方において、単純に目標韻律中のピッチやパワーに合うように接続音声素片のピッチやパワーを合わせようとすると、目標韻律中のピッチやパワーの細かい変動が接続音声素片のピッチやパワーに影響を及ぼしてしまい、逆に音質劣化が謙虚になってしまう。

【００４０】

そこで、本実施形態では、目標韻律中のピッチやパワーの大局的な変動が発話者の抑揚すなわち感情を良く表していると考え、目標韻律からピッチやパワーの緩やかな変動を抽出し、その変動データに基づいて接続音声素片のピッチやパワーをシフトさせることにより、目標韻律中に含まれる抑揚情報が良く反映された合成音声を生成する。

【００４１】

そのために、ＣＰＵ３０１は、ステップＳ４０４の波形接続処理の後、ピッチ適応処理を実行する（ステップＳ４０５）。図５は、ピッチ適応処理の説明図である。図５（ａ）に例示されるように、ＣＰＵ３０１はまず、目標韻律からピッチ周波数の時間変化をピッチ列として抽出する。次に、図５（ｂ）に例示されるように、ＣＰＵ３０１は、ピッチ列の各周波数値を適当な粗さで量子化し、量子化されたピッチ列を算出する。これにより、目標韻律中の微細なピッチの変動が排除され、ピッチの変化の概形が得られる。さらに、図５（ｃ）に例示されるように、ＣＰＵ３０１は、量子化されたピッチ列に対して、時間方向の加重移動平均を演算することによって時間方向の平滑化を行い、平滑化されたピッ

チ列を算出する。具体的には例えば、ＣＰＵ３０１は、量子化されたピッチ列で演算中心サンプル位置を先頭から１サンプルずつ移動させながら、その演算中心サンプル位置の両側所定サンプル分について、例えば演算中心サンプル位置から遠ざかるに従って周波数値が一定量ずつ線形に小さくなるようにして、それらの平均値を算出し、その平均値をその演算中心サンプル位置の演算された値として算出する。これにより、図５（ａ）に例示される細かく変動するピッチ列に対応して、図５（ｃ）に例示されるような自然なピッチ変化を有する平滑化されたピッチ列を得ることができる。ＣＰＵ３０１は、このようにして生成した平滑化されたピッチ列の各時間ごとのピッチの値対応するように、ステップＳ４０４で出力された接続音声素片の各時間ごとのピッチをシフトさせ、その結果を出力する。

【００４２】

続いて、ＣＰＵ３０１は、ステップＳ４０５のピッチ適応処理の後、パワー適応処理を実行する（ステップＳ４０６）。なお、ピッチ適応処理とパワー適応処理の実行順番はどちらでもよく、また、どちらか一方のみが実行されてもよい。図６は、パワー適応処理の説明図である。ＣＰＵ３０１はまず、図６（ａ－１）に例示されるように、目標韻律からパワー値の列（以下「パワー列」）を抽出し、同様に、図６（ａ－２）に例示されるように、接続音声素片（ステップＳＳ４０５のピッチシフトの結果）からパワー列を抽出する。次に、ＣＰＵ３０１は、それぞれのパワー列に対して、ピッチ列の場合と同様にして時間方向の加重移動平均を演算することにより時間方向の平滑化を行い、図６（ｂ－１）に例示される目標韻律に対応する平滑化されたパワー列と、図６（ｂ－２）に例示される接続音声素片に対応する平滑化されたパワー列を算出する。これにより、それぞれのパワー列において、微細な変動が排除され、パワーの変化の概形が得られる。さらに、ＣＰＵ３０１は、目標韻律に対応する平滑化されたパワー列の各時間ごとのサンプル値と、図６（ｂ－２）に例示される接続音声素片に対応する平滑化されたパワー列の各時間ごとのサンプル値との比を算出する。そして、ＣＰＵ３０１は、各時間ごとに算出した比の値を、接続音声素片（ステップＳ４０５のピッチシフトの結果）の各サンプル値に乗算し、その結果を最終的な合成音声として出力する。

【００４３】

ＣＰＵ３０１は、上述のようにして出力された合成音声データを、例えばＲＡＭ３０３や外部記憶装置３０６に音声ファイルとして保存するとともに、図１の音声出力装置１１４を介して合成音声として放音させる。

【００４４】

図７は、図４のステップＳ４０５のピッチ適応処理の詳細例を示すフローチャートである。

【００４５】

ＣＰＵ３０１はまず、図４のステップＳ４０２で生成された目標韻律からピッチ列（以下これを「目標ピッチ列」と記載）を抽出し、この目標ピッチ列と接続音声素片のピッチ列の時間スケールを合わせるタイムストレッチ処理を実行する（ステップＳ７０１）。これにより、両者の時間の長さの違いが吸収される。

【００４６】

次に、ＣＰＵ３０１は、ステップＳ７０１でタイムストレッチ処理した目標ピッチ列と接続音声素片のピッチ列のピッチ存在区間を調整する（ステップＳ７０２）。具体的には、ＣＰＵ３０１は例えば、接続音声素片のピッチ列と目標ピッチ列とを比較し、接続音声素片でピッチの存在しない区間に対応する目標ピッチ列の区間のピッチを削除する。

【００４７】

次に、ＣＰＵ３０１は、ステップＳ７０２でピッチ存在区間を調整した後の目標ピッチ列の周波数値を量子化する（図５（ｂ）に対応）（ステップＳ７０３）。具体的には、ＣＰＵ３０１は例えば、ピッチ周波数値を１オクターブあたりＮ分割（より具体的には３～１０分割等）した単位で、目標ピッチ列を量子化する。

【００４８】

さらに、ＣＰＵ３０１は、ステップＳ７０３で量子化した目標ピッチ列を、図５（ｃ）で前述した加重移動平均演算によって平滑化する（ステップＳ７０４）。

【００４９】

最後に、ＣＰＵ３０１は、ステップＳ７０４で算出された平滑化された目標ピッチ列を、接続音声素片に適応させる（ステップＳ７０５）。具体的には、図５で前述したように、ＣＰＵ３０１は、ステップＳ７０４で平滑化されたピッチ列の各時間ごとのピッチの値対応するように、ステップＳ７０１で調整された接続音声素片の各時間ごとのピッチをシフトさせ、その結果を出力する。

【００５０】

図８は、図４のステップＳ４０６のパワー適応処理の詳細例を示すフローチャートである。

【００５１】

ＣＰＵ３０１はまず、図４のステップＳ４０２で生成された目標韻律からパワー列（以下これを「目標パワー列」と記載）を抽出し、この目標パワー列と接続音声素片のパワー列の時間スケールを合わせるタイムストレッチ処理を実行する（ステップＳ８０１）。なお、図７のステップＳ７０１で実行されたタイムストレッチ処理の結果にスケールが合うように調整される。

【００５２】

次に、ＣＰＵ３０１は、ステップＳ８０１でタイムストレッチ処理した目標パワー列と接続音声素片のパワー列のそれぞれを、図６（ｂ－１）および（ｂ－２）で前述した加重移動平均演算によって平滑化する（ステップＳ８０２）。

【００５３】

続いて、ＣＰＵ３０１は、ステップＳ８０２で算出された目標韻律に対応する平滑化されたパワー列の各時間ごとのサンプル値と、接続音声素片に対応する平滑化されたパワー列の各時間ごとのサンプル値との比を算出する（ステップＳ８０３）。

【００５４】

最後に、ＣＰＵ３０１は、ステップＳ８０３で各時間ごとに算出した比の値を、接続音声素片に適応させる（ステップＳ８０４）。具体的には、図６で前述したように、ＣＰＵ３０１は、ステップＳ８０３で各時間ごとに算出した比の値を、接続音声素片の各サンプル値に乗算し、その結果を最終的な合成音声として出力する。

【００５５】

以上説明した実施形態では、目標韻律中のピッチやパワーの大局的な変動が発話者の抑揚すなわち感情を良く表していると考え、目標韻律からピッチやパワーの緩やかな変動を抽出し、その変動データに基づいて接続音声素片のピッチやパワーをシフトさせることにより、目標韻律中に含まれる抑揚情報が良く反映された合成音声を生成している。しかし本発明では、抑揚情報は目標韻律中のピッチやパワーの大局的な変動に限られるものではない。例えば、抑揚情報として、図４のステップＳ４０１で音素列とともに抽出されるアクセント情報を用い、アクセント位置で図４のステップＳ４０４の波形接続処理で出力される接続音声素片に対して何らかの加工を行うような適応処理が実行されてもよい。その他、抑揚情報を表現できるようなパラメータを入力音声データから抽出することができれば、そのパラメータによって接続音声素片を加工するような適応処理が実行されてもよい。

【００５６】

以上のようにして、本実施形態では、波形接続方式の音声合成システムにおいて目標韻律を音声入力によって指定する際に、音声入力による抑揚指定の高い自由度を維持し、かつ音声コーパスの規模を拡大させる必要なく、合成音声の音質を向上させることが可能となる。

【００５７】

以上の実施形態に関して、更に以下の付記を開示する。

（付記１）

入力されたテキストデータに基づいて、音声素片を選択する選択部と、
前記選択された音声素片を接続することにより接続音声素片を生成する接続部と、
前記入力された音声データに含まれる韻律情報から抑揚情報を抽出する抑揚情報抽出部と、
前記接続音声素片に含まれる要素の少なくとも一部を前記抑揚情報に基づいて変更して合成音声を出力する出力部と、
を備える音声合成装置。
（付記２）
前記出力部は、前記接続音声素片に含まれる要素であるピッチ列を、前記抑揚情報抽出部から抽出された抑揚情報に適応するように変更するピッチ適応部を含む、付記１に記載の音声合成装置。
（付記３）
前記抑揚情報抽出部は、前記入力された音声データに含まれる韻律情報としてのピッチ列のピッチを平滑化し、当該ピッチの平滑化されたピッチ列を前記抑揚情報として抽出する、付記２に記載の音声合成装置。
（付記４）
前記抑揚情報抽出部は、前記入力された音声データに含まれる韻律情報としてのピッチ列を構成するピッチを量子化し、当該量子化されたピッチを加重移動平均演算することにより、前記ピッチの平滑化されたピッチ列を生成する、付記３に記載の音声合成装置。
（付記５）
前記ピッチ適応部はさらに、前記入された力音声データに含まれる韻律情報としてのピッチ列と前記接続音声素片に含まれるピッチ列との時間スケールを調整するとともに、前記韻律情報としてのピッチ列と前記接続音声素片に含まれるピッチ列とのピッチ存在区間を調整する、付記２ないし４のいずれかに記載の音声合成装置。
（付記６）
前記出力部は、前記接続音声素片に含まれる要素であるパワー列を、前記抑揚情報抽出部から抽出された抑揚情報に適応するように変更するパワー適応部を含む、付記１に記載の音声合成装置。
（付記７）
前記抑揚情報抽出部は、前記入された力音声データに含まれる韻律情報としてのパワー列平滑化し、当該平滑化されたパワー列を前記抑揚情報として抽出し、
前記パワー適応部は、前記接続音声素片に含まれるパワー列を平滑化し、当該平滑化されたパワー列と前記抑揚情報としての平滑化されたパワー列との比の列を算出し、当該比の列に基づいて前記接続音声素片のパワー列を修正する、付記６に記載の音声合成装置。
（付記８）
前記前記平滑化されたパワー列は、前記パワー列に含まれるパワーそれぞれを加重平均演算することにより取得する、付記７に記載の音声合成装置。
（付記９）
前記パワー適応部はさらに、前記入力された音声データに含まれる韻律情報としてのパワー列及び前記接続音声素片に含まれるパワー列それぞれの時間スケールを調整する、付記６ないし８のいずれかに記載の音声合成装置。
（付記１０）
音声合成装置に用いられる音声合成方法であって、前記音声合成装置が、
入力されたテキストデータに基づいて、音声素片を選択し、
前記選択された音声素片を接続することにより接続音声素片を生成し、
前記入力された音声データに含まれる韻律情報から抑揚情報を抽出し、
前記接続音声素片に含まれる要素の少なくとも一部を前記抑揚情報に基づいて変更して合成音声を出力する、音声合成方法。
（付記１１）
音声合成装置に用いられる音声合成装置として用いられるコンピュータに、

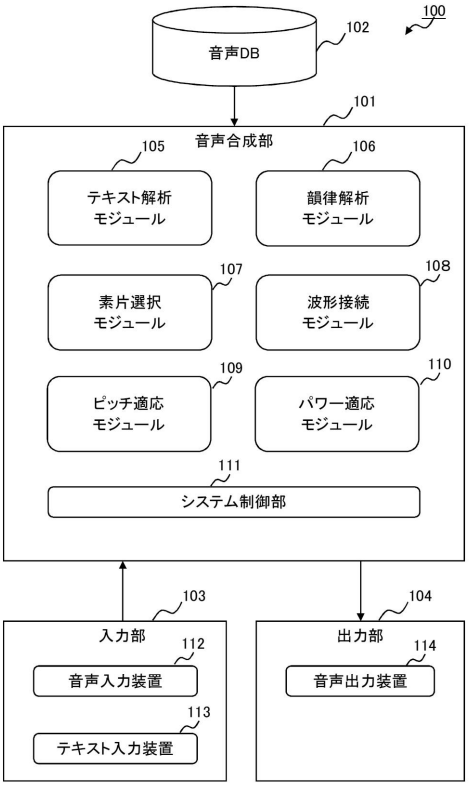
入力されたテキストデータに基づいて、音声素片を選択するステップと、
前記選択された音声素片を接続することにより接続音声素片を生成するステップと、
前記入力された音声データに含まれる韻律情報から抑揚情報を抽出するステップと、
前記接続音声素片に含まれる要素の少なくとも一部を前記抑揚情報に基づいて変更して合成音声を出力するステップと、
を実行させるプログラム。

【符号の説明】

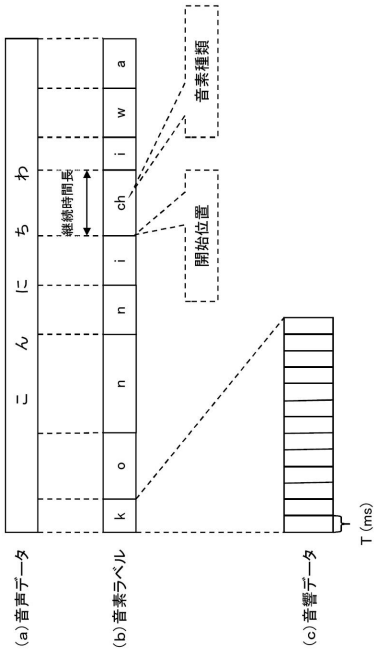
【００５８】

１００　音声合成装置
１０１　音声合成部
１０２　音声ＤＢ（データベース）
１０３　入力部
１０４　出力部
１０５　テキスト解析モジュール
１０６　韻律解析モジュール
１０７　素片選択モジュール
１０８　波形接続モジュール
１０９　ピッチ適応モジュール
１１０　パワー適応モジュール
１１１　システム制御部
１１２　音声入力装置
１１３　テキスト入力装置
１１４　音声出力装置
３０１　ＣＰＵ
３０２　ＲＯＭ
３０３　ＲＡＭ
３０４　入力装置
３０５　出力装置
３０６　外部記憶装置
３０７　可搬記録媒体駆動装置
３０８　通信インタフェース
３０９　バス
３１０　可搬記録媒体

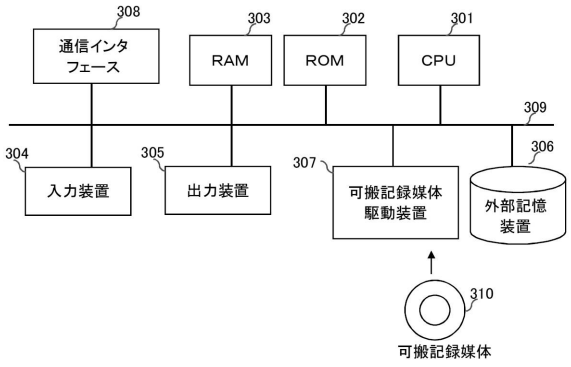
【図１】


100
音声DB
102
101
音声合成部
105
テキスト解析
モジュール
106
韻律解析
モジュール
107
素片選択
モジュール
108
波形接続
モジュール
109
ピッチ適応
モジュール
110
パワー適応
モジュール
111
システム制御部
103
入力部
112
音声入力装置
113
テキスト入力装置
104
出力部
114
音声出力装置

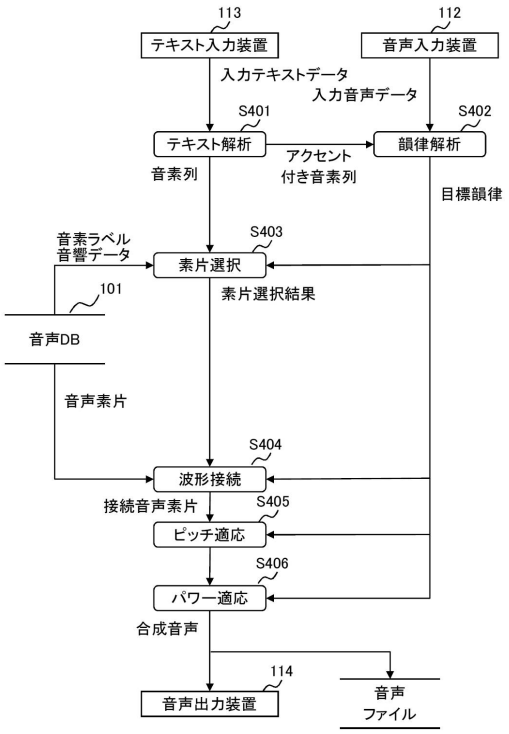
【図２】


こんにちわ
k o n n i ch i w a
継続時間長
音素種類
開始位置
T (ms)
(a)音声データ
(b)音素ラベル
(c)音響データ

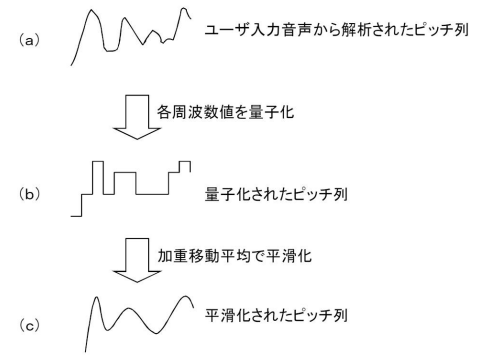
【図３】


308
通信インタ
フェース
303
RAM
302
ROM
301
CPU
309
304
入力装置
305
出力装置
307
可搬記録媒体
駆動装置
306
外部記憶
装置
310
可搬記録媒体

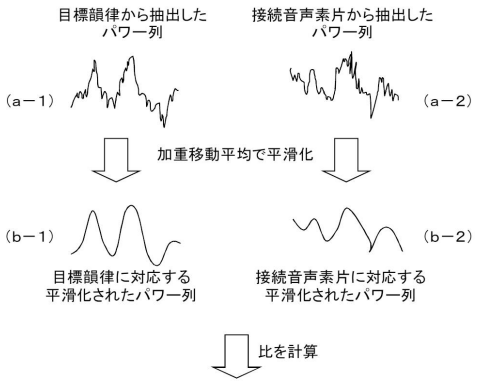
【図４】


113
テキスト入力装置
112
音声入力装置
入力テキストデータ
入力音声データ
S401
テキスト解析
S402
韻律解析
音素列
アクセント
付き音素列
目標韻律
音素ラベル
音響データ
S403
素片選択
素片選択結果
101
音声DB
音声素片
S404
波形接続
接続音声素片
S405
ピッチ適応
S406
パワー適応
合成音声
114
音声出力装置
音声
ファイル

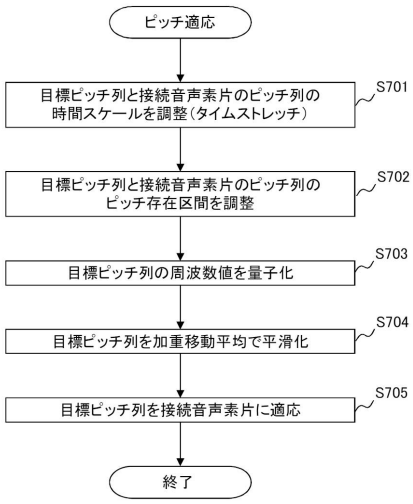
【図５】


(a)
ユーザ入力音声から解析されたピッチ列
各周波数値を量子化
(b)
量子化されたピッチ列
加重移動平均で平滑化
(c)
平滑化されたピッチ列

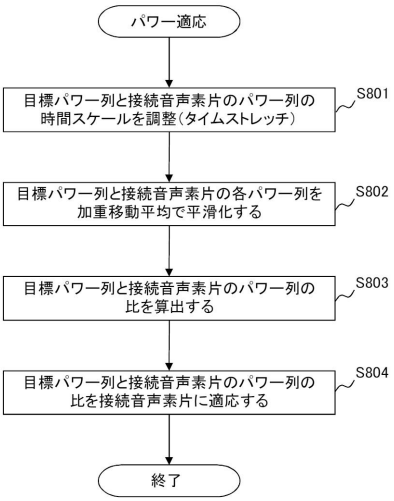
【図６】


目標韻律から抽出した
パワー列
接続音声素片から抽出した
パワー列
(a－1)
(a－2)
加重移動平均で平滑化
(b－1)
(b－2)
目標韻律に対応する
平滑化されたパワー列
接続音声素片に対応する
平滑化されたパワー列
比を計算

【図７】


ピッチ適応
目標ピッチ列と接続音声素片のピッチ列の
時間スケールを調整（タイムストレッチ）
S701
目標ピッチ列と接続音声素片のピッチ列の
ピッチ存在区間を調整
S702
目標ピッチ列の周波数値を量子化
S703
目標ピッチ列を加重移動平均で平滑化
S704
目標ピッチ列を接続音声素片に適応
S705
終了

【図８】


パワー適応
目標パワー列と接続音声素片のパワー列の
時間スケールを調整（タイムストレッチ）
S801
目標パワー列と接続音声素片の各パワー列を
加重移動平均で平滑化する
S802
目標パワー列と接続音声素片のパワー列の
比を算出する
S803
目標パワー列と接続音声素片のパワー列の
比を接続音声素片に適応する
S804
終了

フロントページの続き

(56)参考文献 特開２０１２－２２０７０１（ＪＰ，Ａ）
特開２０００－０１０５８１（ＪＰ，Ａ）
特開２００６－３０９１６２（ＪＰ，Ａ）
特開平１０－１５３９９８（ＪＰ，Ａ）
米国特許第０５９４０７９７（ＵＳ，Ａ）
特開２０１０－００９０３４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１３／００－１３／１０